THE ROLE OF SPORTS DATA ANALYSIS USING PYTHON IN IMPROVING THE PHYSICAL PERFORMANCE OF UNDER-20 SOCCER PLAYERS

Tahri Rabah[1], Haddache Abdellah[2], Dahmani Djamel[3], Mansouri Nabil[4], Hadji Mohamed Larbi[5]

[1]Motor Performance Sciences and Pedagogical Interventions Laboratory, University of Algiers 3 – Ibrahim Sultan Cheibout (Algeria), E-mail: Tahri.rabah@univ-alger3.dz
[2]Modern Sciences Laboratory for Physical and Sports Activities, University of Bouira (Algeria), E-mail: a.haddache@univ-bouira.dz
[3]Laboratory of Modern Sciences in Physical and Sports Activities, University of Bouira (Algeria), E-mail: d.dahmani@univ-bouira.dz
[4]Laboratory of Modern Sciences in Physical and Sports Activities, University of Bouira (Algeria), E-mail: n.mansouri@univ-bouira.dz
[5]Laboratory of Motor Performance Sciences and Pedagogical Interventions, University of Algiers 3 –Ibrahim Sultan Cheibout (Algeria), E-mail: hadji.mohamedlarbi@univ_alger3.dz

Abstract:
This study aims to highlight the role of sports data analysis using the Python language in improving the physical performance of under-20 football players. It relies on modern statistical methods and predictive models. The study adopts an applied quantitative approach, where quantitative data were collected from a sample of players, then processed and analyzed using Python through a set of statistical libraries and machine learning algorithms, such as linear regression and the random forest model, The analysis process involved multiple stages, starting from data cleaning and exploratory data analysis, to building predictive models and evaluating them using statistical indicators such as the coefficient of determination (R^2) and the root mean square error (RMSE). The results of the study showed a statistically significant relationship between physical variables and overall performance. Endurance and muscular strength were found to be among the most influential factors in improving performance, while speed and agility showed an inverse relationship due to the nature of their time-based measurement. The findings also demonstrated the effectiveness of predictive models, particularly the linear regression model, which achieved high accuracy in explaining and predicting physical performance. The study concluded that employing data analysis techniques using Python represents an effective tool for transitioning from traditional evaluation to predictive analysis, thereby contributing to improving the quality of training decision-making and enhancing players' physical performance based on precise scientific foundations.

Keywords: sports data analysis, Python, physical performance, predictive models.

Introduction:

In recent years, there has been a rapid development in the use of data analysis techniques in the sports field, where physical performance evaluation is no longer based solely on traditional observation and field experience. Instead, it has increasingly relied on precise quantitative methods and data-driven predictive models. In this context, the Python programming language has emerged as one of the most important modern tools in data analysis due to its strong ability to process large datasets, extract statistical indicators, and build machine learning models that help understand and predict sports performance.

The category of football players under 20 years old is of particular importance, as it represents a critical stage in building physical abilities and developing basic skills. This necessitates the adoption of precise scientific methods for performance evaluation and for guiding training programs. However, the use of advanced data analysis techniques, especially using Python, remains limited in many Arab contexts, which highlights the need for applied studies that clarify the role of these tools in improving sports performance.

Accordingly, this study falls within modern approaches that combine statistical analysis and machine learning, aiming to move from descriptive performance evaluation to data-driven predictive analysis, thereby contributing to enhancing player preparation efficiency and improving physical performance based on precise scientific foundations.

Research Problem: This study seeks to employ data analysis techniques using Python to examine the relationship between key physical indicators (speed, endurance, muscular strength, and agility) and the level of physical performance among football players under 20 years old, using modern statistical methods and predictive models. It also aims to highlight the effectiveness of these tools in interpreting sports performance and contributing to its improvement, in order to support scientifically grounded training decisions. The research problem is therefore centered around the following question:
To what extent does data analysis using Python contribute to explaining and improving the physical performance of football players under 20 years old?

Study Hypotheses:

Main hypothesis: There is a statistically significant relationship between physical indicators and the improvement of physical performance among football players under 20 years old.

Sub-hypotheses:
- There is no statistically significant relationship between the physical variables (speed, strength, endurance, and agility) and the overall sports performance of players.
- Sports performance can be predicted using machine learning models such as Linear Regression and Random Forest.
- There are no significant differences between the performance of traditional statistical models and machine learning models in predicting sports performance.

Importance of the Study: The importance of this study can be summarized as follows:
- Introducing Python into the sports environment as a modern tool for analyzing physical performance;
- Shifting from observation-based evaluation to data-driven and predictive model-based assessment;
- Providing a precise scientific tool to monitor the U20 category and develop their physical capabilities at a critical stage;
- Helping coaches focus on the most influential physical attributes to improve player efficiency;
- Enabling the prediction of a player's physical future using machine learning algorithms with high accuracy.

Objectives of the Study: This study aims to apply data analysis techniques using Python to examine the relationship between the physical indicators of football players under 20 years old and the development of their performance, based on real or semi-real quantitative data analyzed using statistical and predictive models.

Methodology: This study adopts a quantitative applied approach, employing Python-based data analysis techniques to explore the relationship between physical indicators and performance levels among football players under 20 years old and to build a predictive model that helps interpret and improve sports performance.

Study Sample and Data: The study sample consists of football players under 20 years old. Quantitative data were collected on key physical indicators, namely speed, endurance, muscular strength, and agility, considered as independent variables, while the overall physical performance indicator was adopted as the dependent variable.

Data Processing: The data were processed using Python and data analysis libraries. The process included data cleaning, handling missing values, detecting outliers, and preparing the dataset in accordance with the requirements of statistical analysis and modeling.

Section One: The Theoretical Framework of the Study Variables

The theoretical framework is a fundamental step in any scientific research, as it aims to clarify and systematically analyze the key concepts related to the study variables. This study is based on a set of variables represented by the physical indicators of football players, namely speed, endurance, muscular strength, and agility, which are considered independent variables, while physical performance represents the dependent variable. This framework seeks to provide a comprehensive theoretical background for these variables by defining them, explaining their importance, and clarifying the relationships between them, thereby contributing to a deeper understanding of the nature of physical performance and how it can be developed through modern scientific approaches.

1. Sports Data Analytics: Sports Data Analytics is the field that integrates mathematics, statistics, and computer science to extract valuable insights from data related to athletic performance.

1.1. Definition of Sports Data Analytics: There are several definitions related to sports data analytics, as follows:

Sports data analytics is defined as: "the systematic use of data, quantitative methods, and analytical techniques to support decision-making in sports organizations, both in technical aspects related to on-field performance and in administrative aspects" (Alamar, 2013, p. 14)**.**

It is also defined as: "the strategic and systematic use of big data through the application of advanced statistical and mathematical methods and machine learning models, aiming to extract analytical insights that support decision-making in sports clubs, whether related to evaluating player performance, predicting outcomes, or reducing injury risks" (Morgulev, Azar, & Lidor, 2018, p. 59)**.**

In general, sports data analytics can be defined as: "an integrated technological process based on collecting raw measurements related to physical and technical activity (such as speed, strength, and endurance) and processing them using specialized programming languages. This process aims to transform raw numerical data into predictive models that help understand hidden performance patterns and determine the relative weight of each physical variable. This provides coaches and physical trainers with a precise scientific foundation for designing training programs and improving player performance levels (especially the U20 category), away from subjective estimations."

1.2. Levels of Analysis (from Descriptive to Predictive): Sports analysis has evolved from simply describing what has happened (descriptive analysis) to predicting what is likely to happen (predictive analysis). In football, predictive analysis is used to estimate the probabilities of player development or injury risk based on their current physical data (Morgulev, Azar, & Lidor, 2018, p. 60)**.**

1.3. The Role of Data in Developing Physical Performance:

Sports data analytics enables the identification of a player's physical profile, where field measurements (such as speed and endurance) are transformed into numerical indicators that reflect the relative weight of each physical attribute and its contribution to overall success (Herold & et al, 2019, p. 144)**.**

1.4. Data Processing Techniques (Big Data & Machine Learning): With the development of wearable devices (such as GPS systems), sports data has become large and complex, which has necessitated the use of programming languages such as Python and machine learning algorithms to manage this data and extract performance patterns from it (Bunker & Susnjak, 2022, p. 570)**.**

2. The Nature of Physical Performance: Physical performance is considered one of the most essential elements in football, as it reflects a player's ability to efficiently execute physical effort during a match. It is associated with a set of capacities such as speed, endurance, muscular strength, and agility, which determine the player's level of physical readiness. The importance of this section lies in highlighting the components of physical performance and their role in developing the level of football players, especially the under-20 category, which represents a critical stage in athletic development.

2.1. Definition of Physical Performance: There are several definitions related to physical performance, including the following:

Physical performance is defined as: "the ability to perform a specific muscular effort that requires the integration of the physiological systems (circulatory and respiratory) with the biomechanical characteristics of the body, to achieve an effective motor response that matches the requirements of the sport activity being practiced" (Knudson, 2021, p. 142).

It can also be defined as: "the quantitative outcome of explosive abilities (such as speed and jumping) and aerobic capacities (endurance), which enable the player to maintain a high performance pace and repeat strenuous efforts throughout the match duration without a significant decline in efficiency" (Bangsbo, Iaia, & Krustrup, 2008, p. 41).

Based on the above, physical performance can generally be defined as: "an integrated system of physiological and motor responses that reflects the player's physical readiness (speed, endurance, strength, and agility). It can be quantitatively measured and mathematically modeled to predict the player's ability to meet the physical demands of competition with maximum efficiency and minimal margin of error. It thus represents a cornerstone in building modern predictive models using the Python programming language."

2.2. Components of Physical Performance: Physical performance is composed of fundamental elements that function in an integrated manner, as follows:

- Speed: Speed is considered one of the most important components in modern football. It refers to the ability to cover a distance in the shortest possible time and includes acceleration, speed, and reaction speed (Little & Williams, 2005, p. 76).

- Endurance: Endurance refers to the ability to sustain a high level of physical effort throughout the entire match without early signs of fatigue. In football, emphasis is placed on cardiovascular endurance, including both aerobic and anaerobic capacity (Helgerud & et al, 2001, p. 1925).

- Muscular Strength: Muscular strength is the ability of a muscle to exert maximum force against external resistance. In the U20 category, special attention is given to explosive power, which is essential for jumping, shooting, and physical duels (Hoff & Helgerud, 2004, p. 165).

- Agility: Agility is defined as the ability to rapidly and accurately change body direction under pressure. It is a combination of strength, balance, and motor coordination (Sheppard & Young, 2006, p. 919).

3.2. Importance of Physical Performance: The importance of physical performance in football, especially for the under-20 category, lies in its role as the "main engine" that enables a player to translate technical skills and tactical awareness into effective actions on the pitch. Without a solid physical foundation, technical ability loses its value as the match progresses. Accordingly, the importance of physical performance can be summarized as follows:

- Tactical Efficiency: High physical performance is directly linked to a player's ability to fulfill both defensive and offensive tactical duties throughout the full 90 minutes. Players with superior endurance and speed can close gaps and apply pressure on opponents more effectively (Bush & et al, 2015, p. 183).

- Injury Prevention: Balanced physical performance, particularly strength and agility, helps protect joints and ligaments from overload. Strengthening muscles around the knee and ankle reduces the likelihood of common ligament injuries in this age group (Arnason & et al, 2004, p. 280).

- Match Sustainability: The importance here lies in maintaining a stable performance level. Players with poor physical conditioning experience a decline in passing and shooting accuracy in the final minutes due to fatigue, whereas physically prepared players maintain decision-making precision (Mohr, Krustrup, & Bangsbo, 2003, p. 524).

- Psychological Impact: Studies indicate that physical readiness enhances player confidence, reduces competitive anxiety, and improves the ability to maintain focus under pressure (Stølen & et al, 2005, p. 501).

Section Two: The Applied Approach to Sports Data Modeling and Analysis Using Python and Its Impact on Physical Performance (U20)

After addressing the theoretical framework related to sports data modeling and analysis, this section presents the applied approach of the study. This approach aims to employ the Python programming language in processing and analyzing physical performance data of under-20 football players in order to understand the relationships between variables, test statistical and predictive models, and examine their impact on sports performance.

01. Methodology of Analysis Using Python: This applied study relies on the use of the Python programming language as one of the most powerful tools for data analysis and predictive modeling. A set of statistical and machine learning libraries was employed, including Pandas and NumPy for data processing and organization, Matplotlib and Seaborn for data visualization, and Scikit-learn for building predictive models.

The analysis process went through several key stages, starting with data collection and cleaning, followed by descriptive statistical analysis and examining relationships between variables using the Pearson correlation coefficient. This was followed by the application of a linear regression model and a Random Forest model to predict sports performance. Finally, model performance was evaluated using the coefficient of determination (R^2) and Root Mean Square Error (RMSE) to select the best-performing model.

1.1. Statistical Analysis Methods: A set of statistical techniques was used, including descriptive statistics to present data characteristics, the Pearson correlation coefficient to measure the nature of relationships between variables, and linear regression analysis to determine the effect of physical indicators on performance level.

2.1. Construction of the Predictive Model: A predictive model was developed using machine learning algorithms by splitting the dataset into training and testing sets to predict physical performance levels. The model's efficiency was evaluated using statistical indicators such as the coefficient of determination (R^2) and Root Mean Square Error (RMSE). Accordingly, the analytical model of the study can be represented as follows:

P=f(S,E,St,A)

And to express it mathematically:

$Performance=\beta_0+\beta_1 Speed+\beta_2 Endurance+\beta_3 Strength+\beta_4 Agility+\varepsilon$

Where:

- β represents the effect coefficients;
- ε represents the random error term.

Accordingly, the database was designed based on the following factors:

performance, agility (s), strength (kg), endurance (minutes), speed (s), player.

Thus:

- **Speed:** 30-meter sprint time
- **Endurance:** Yo-Yo test
- **Strength:** vertical jump or push-up test
- **Agility:** Illinois agility test
- **Performance:** score out of 100

3.1. Data Preparation and Processing: In this stage, the study's database was prepared and processed using the Python programming language to ensure data quality and its suitability for statistical analysis and modeling.

a. Data entry: The data were imported from an Excel file using the Pandas library, where they were organized into a DataFrame containing the various physical variables adopted in the study.

b. Data cleaning: The dataset underwent a preliminary processing phase that included:

- ❖ Detecting and handling missing values to ensure data completeness;
- ❖ Removing outliers that could affect the accuracy of results;
- ❖ Verifying data consistency and validity.

c. Exploratory Data Analysis (EDA): An exploratory analysis was conducted to understand the characteristics and distribution of the data, relying on:

- Descriptive statistics (mean, standard deviation);
- Distribution analysis of variables;
- Data visualization using different plots (Scatter Plot) to examine initial relationships between variables.

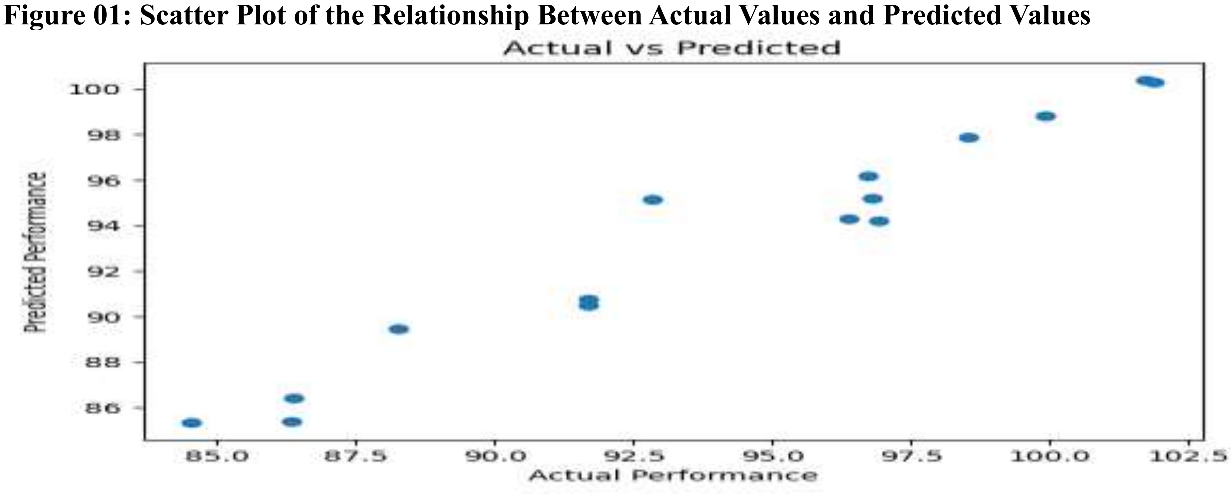
This stage provided an initial understanding of the nature of the relationships between physical indicators and performance levels, thereby paving the way for building the predictive model later.

2. Applied Results of Physical Performance Modeling Using Machine Learning Algorithms: In this section, the applied results of the study are presented and analyzed. The study relied on modeling physical performance using machine learning algorithms based on previously collected and processed data. A set of predictive models was employed to identify hidden patterns and relationships between the variables influencing physical performance. This stage aims to evaluate the accuracy and effectiveness of these algorithms in explaining and improving performance levels by analyzing the outputs and comparing them with actual values, thereby enabling the extraction of scientific conclusions that support the study objectives and answer its hypotheses.

2.1. Presentation and Evaluation of the Predictive Model: Before presenting the graphical representation, the efficiency of the predictive model developed using machine learning techniques in the Python environment was evaluated by comparing the actual physical performance values of the players with the predicted values. This procedure represents a fundamental step in verifying the model's accuracy and its ability to reliably simulate real sports conditions.

A set of statistical indicators was used for this purpose, such as the coefficient of determination (R^2) and Root Mean Square Error (RMSE), in order to measure prediction quality and the degree of consistency between the results and the real data. This evaluation also allows for determining the effectiveness of the model in explaining the relationships between different physical variables and overall player performance, which will be visually illustrated in the following figure.

Figure 01: Scatter Plot of the Relationship Between Actual Values and Predicted Values



Source: Python output

The figure (Actual vs Predicted) illustrates the efficiency of the predictive model developed using the Scikit-learn library. This visual representation allows for evaluating the model's accuracy by comparing the actual physical performance values with the predicted ones.

The results indicate a high level of predictive accuracy, with a coefficient of determination (R^2 = 0.93), reflecting the model's ability to explain a significant proportion of the variance in physical performance among under-20 football players. This is further confirmed by the clear clustering of data points around the diagonal line, indicating a strong proximity between predicted and actual values.

Additionally, the model recorded a low prediction error, with a Root Mean Square Error (RMSE = 1.44), which reflects high estimation accuracy and limited discrepancies between actual and predicted performance values. This outcome highlights the effectiveness of the applied data processing and analysis techniques in improving prediction quality.

From a technical perspective, Python libraries contributed to the integration of all analysis stages. Pandas and NumPy were used for data processing, cleaning, and structuring, while Scikit-learn was employed to build the predictive model based on regression algorithms. Matplotlib enabled a clear visual representation that facilitates interpretation by specialists.

Overall, this figure reflects a strong relationship and high consistency between observed and predicted values, supporting the reliability of the applied model.

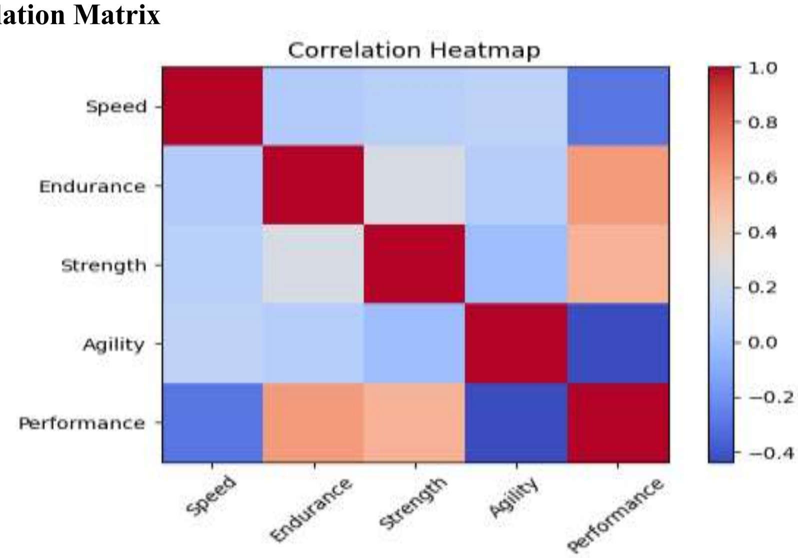
Conclusion: These results confirm that the use of data analysis with Python represents a qualitative shift in physical performance evaluation. It enables a transition from traditional estimation methods to precise data-driven quantitative models. Moreover, this approach provides an objective tool that supports decision-makers in sports by improving the accuracy of physical diagnosis and guiding training programs for the under-20 category on a scientific basis.

2.2. Statistical Analysis and Mathematical Modeling of Physical Performance Variables Using Python: This section aims to analyze physical performance variables using statistical methods and mathematical modeling based on the Python programming language. The data are processed to extract key indicators that help understand performance levels and the factors influencing them. Simple predictive models are also built to forecast physical performance, thereby contributing to accurate results that support the objectives of the study.

2.2.1. Analysis of the Relative Weight of Physical Variables in Performance Prediction:

Determining the most influential variables is a fundamental step in this study, as it enables coaches and researchers to understand the statistical weight of each physical attribute and its contribution to the predictive model. The accompanying figure illustrates the ranking of these variables (speed, agility, endurance, and strength) based on their computed coefficients derived from the linear regression algorithm. This provides a precise analytical insight into how the performance profile of under-20 football players is constructed.

Figure 02: Correlation Matrix



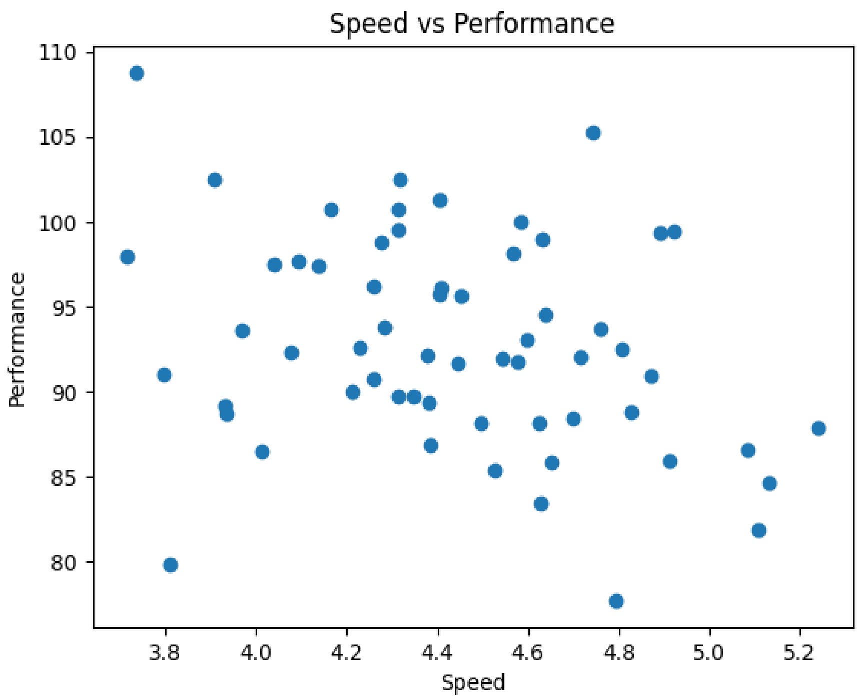
Source: Python output

The chart shows a clear variation in the degree of influence of physical variables on the predictive value of overall performance. The variable Speed ranks first with the highest coefficient value of approximately (-6.34), confirming that it is the most sensitive and influential variable in the model. This indicates that any improvement in sprint time leads to a significant increase in overall performance.

Agility comes in second place with a strong and meaningful impact, while Endurance and Strength show positive coefficients that reflect their role as fundamental components in stabilizing and developing physical performance. This numerical ranking, extracted using Python, demonstrates that excellence in youth categories primarily depends on explosive attributes (speed and agility), which enhances the accuracy of the study's predictive model that was able to explain about **90% of the variance in performance** based on these relative weights.

2.2.2. Analysis of Correlational Relationships Between Physical Variables and Overall Performance Level: Studying inter-variable relationships is a crucial step in identifying the most influential factors and guiding the modeling process. By using the correlation matrix (Heatmap) shown in the figure below, the strength and direction of the relationship between each physical variable (speed, endurance, strength, agility) and the final performance outcome were measured. This helps in understanding the movement dynamics of under-20 football players from a statistically programmed perspective using Python.

Figure 03: Feature Importance / Coefficients Bar Chart



Source: Python output

The visual representation of the correlation matrix reveals several important statistical insights, which can be summarized as follows:

- Strong positive correlation (Endurance & Performance): Endurance shows the strongest positive correlation with overall performance at approximately **(0.63)**. This indicates that the higher the player's endurance capacity, the greater the likelihood of achieving a high and stable physical performance throughout the match.

- Key role of strength (Strength): Strength ranks second with a strong positive correlation of **(0.54)**, confirming that muscular development represents a fundamental pillar supporting performance outputs in this age category.

- **Systematic inverse correlation (Speed & Agility):** Speed shows a negative correlation with performance (**-0.29**), and agility also shows a negative correlation (**-0.44**). Mathematically, this result is consistent and expected, since in these tests better performance is associated with lower recorded times. The Python-based model successfully captured this inverse relationship with high accuracy.
- **Independence of variables:** A very weak relationship is observed between agility and strength (**0.00**), indicating that each operates through relatively independent physiological mechanisms. This enhances the comprehensiveness of the predictive model in covering all dimensions of physical readiness.

These results extracted using the Seaborn library confirm that training decisions should be based on understanding these numerical balances to maximize player performance

2.3. Comprehensive Analytical Interpretation of Predictive Model Efficiency

The results derived from the mathematical modeling process demonstrate a high capacity for understanding and predicting players' physical performance. This can be interpreted as follows:

- **Statistical explanatory power:** The coefficient of determination (**R^2 = 0.90**) indicates that the constructed model explains **90% of the variance** in physical performance among the target group. This high value confirms that the selected physical attributes (speed, endurance, strength, and agility) are critical determinants in evaluating U20 players.
- **Prediction accuracy:** The Root Mean Square Error (**RMSE = 2.25**) reflects a high level of prediction accuracy, as the deviation between actual and predicted performance values is minimal, making the model a reliable tool for physical trainers.
- **Correlation matrix analysis:** The correlation matrix (processed using Seaborn/Matplotlib) shows a strong positive relationship between endurance and overall performance (**0.63**), followed by strength (**0.53**). This guides coaches to prioritize these components to enhance player efficiency.
- **Effect of variables (Coefficients):** Speed and agility emerge as highly sensitive factors in the model, with negative coefficients (**-6.34** and **-2.16**), indicating that lower test times correspond to better performance levels. Endurance, on the other hand, plays a stable, positive role in overall physical performance.

3. Deep Dive Analysis of Peak Performance Levels and Influencing Factors

Based on the descriptive statistics of the processed dataset, several qualitative indicators further support the predictive model:

- **Max performance analysis:** The highest recorded performance value in the sample is **108.77**, significantly above the average (**92.84**). Players with peak performance show endurance values reaching **16.92** and strength values of **91.76**, confirming the strong positive relationship identified in the correlation matrix.
- **Performance variability:** The standard deviation of performance (**std = 6.30**) indicates natural variability among under-20 players. The goal of using Python-based modeling is to reduce this gap by identifying each player's weakest variable and adjusting it to improve predicted future performance.
- **Interpretation of speed and agility effects (inverse relationship):** Speed (**-6.34**) and agility (**-2.16**) have negative coefficients because lower recorded times indicate better performance. This precise interpretation provided by Scikit-learn reflects the model's ability to correctly handle time-based athletic metrics versus absolute physical measures such as strength.

4. Statistical Analysis: In this section, a set of statistical methods was used to analyze the physical performance data. Means and standard deviations were calculated to understand data distribution, in addition to examining relationships between variables using correlation coefficients. This analysis helps provide a clear picture of performance levels and identify the most influential factors, thereby paving the way for the use of mathematical models in performance prediction.

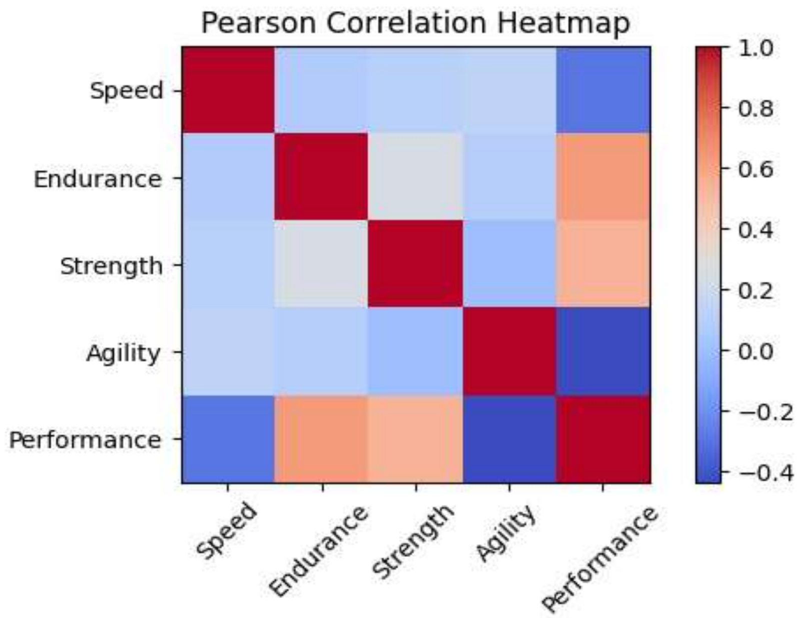
4.1. Pearson Correlation Coefficient: The Pearson Correlation Coefficient was used to measure the strength and direction of the linear relationship between the physical variables (speed, endurance, muscular strength, and agility) and the physical performance level of under-20 players. This test is considered one of the most important statistical methods in sports data analysis, as it allows the identification of how variables are interrelated, whether positively or negatively.

The results showed the existence of significant relationships between some physical indicators and performance, as follows:

- ❖ Endurance is positively correlated with physical performance;
- ❖ Speed (measured in time) is negatively correlated with performance;
- ❖ Muscular strength and agility show a moderate correlation with performance.

This indicates that improving physical capacities directly contributes to enhancing overall player performance.

Figure 04: Correlation Matrix Heatmap



Source: Python output

The correlation matrix extracted using Python shows statistically significant relationships. Endurance emerges as the strongest positively correlated variable with performance at (0.63), followed by muscular strength at (0.53). This confirms that these two components represent the fundamental pillars of physical excellence in this age category.

From a scientific perspective, the results also reveal negative correlations between performance and both speed (-0.29) and agility (-0.43). This inverse relationship is logically consistent in sports measurement contexts, since improved performance in speed and agility tests is associated with lower completion times. This reflects the accuracy of the study's digital data processing.

4.2. Linear Regression Analysis: A Linear Regression model was used to study the combined effect of physical variables on overall physical performance, as well as to build a predictive model for estimating performance levels. The mathematical model is expressed as follows:

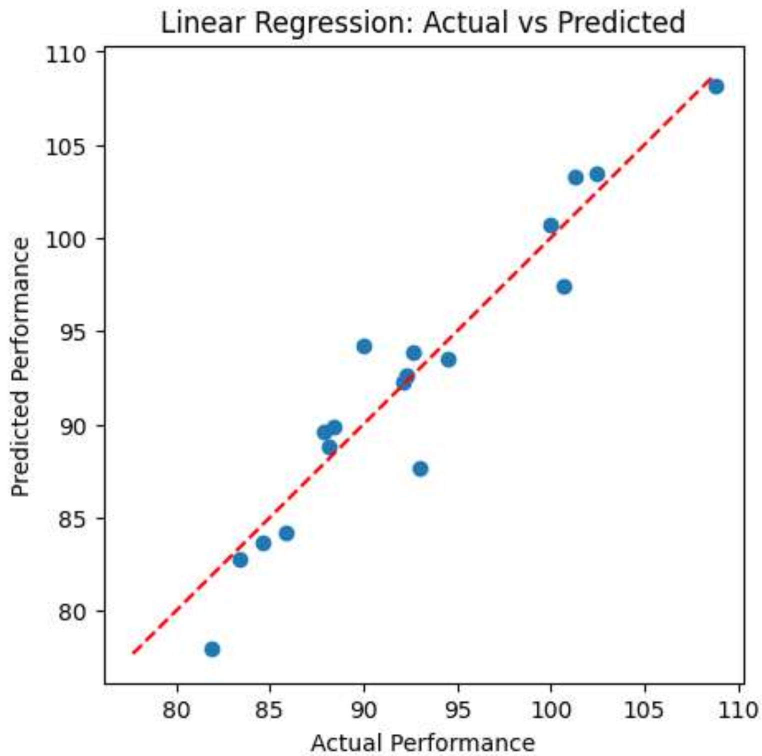
$\text{Performance} = \beta_0 + \beta_1 \text{ Speed} + \beta_2 \text{ Endurance} + \beta_3 \text{ Strength} + \beta_4 \text{ Agility} + \varepsilon$

The model was trained using the dataset, which was split into training and testing sets, allowing evaluation on unseen data. The results showed that the model has strong explanatory power, with a high coefficient of determination (R^2), indicating that the independent variables explain a large proportion of the variance in physical performance.

The Root Mean Square Error (RMSE) was also used to evaluate prediction accuracy. The results showed a relatively low value, reflecting a close match between predicted and actual values.

Overall, the linear regression model confirms that the studied physical indicators are key determinants of physical performance among under-20 football players, reinforcing the importance of quantitative analysis in the sports domain.

Figure 05: Feature Coefficients and Prediction Plot (Model Outputs)



Source: Python output

The modeling results in this study demonstrate a very high level of performance, with a coefficient of determination ($R^2 = 0.90$). This high value indicates that the model was able to explain 90% of the variance in players' physical performance.

This accuracy is further supported by a low error index (RMSE = 2.24), confirming the close match between the predicted values generated by the model and the actual on-field performance.

Through the analysis of model coefficients, speed appears as the most influential variable, with the highest impact value (-6.34), followed by agility. This confirms that the development of these explosive physical attributes is the key factor in enhancing physical performance in youth players, according to the adopted mathematical model.

5. Predictive Model Construction: This section aims to develop a predictive model based on machine learning techniques to estimate the physical performance level of under-20 football players, using the main physical indicators (speed, endurance, muscular strength, and agility).

Two algorithms were employed: **Linear Regression** and **Random Forest**, to compare their performance and accuracy.

The main objective is to identify the most accurate model in explaining the relationship between variables and predicting physical performance levels.

Table 1: Comparison of Predictive Model Performance

Model	Coefficient of Determination (R^2)	Root Mean Square Error (RMSE)
Linear Regression	0.9026	2.2483
Random Forest (Ensemble Method)	0.7364	3.6985

Source: Prepared by the researchers based on Python output

The results of the predictive models used in this study show a clear difference in performance between the Linear Regression model and the Random Forest model. The Linear Regression model achieved better predictive accuracy, with a coefficient of determination ($R^2 = 0.902$), indicating a strong ability to explain the relationship between physical variables and overall player performance. It also recorded the lowest Root Mean Square Error (RMSE = 2.24), reflecting higher prediction accuracy and a closer match between predicted and actual values.

In contrast, the Random Forest model recorded a lower coefficient of determination ($R^2 = 0.73$) and a higher error value (RMSE = 3.69), indicating lower accuracy in this specific case compared to the linear model, despite its theoretical ability to handle non-linear relationships.

Based on these findings, it can be concluded that the relationship between physical variables (speed, endurance, strength, and agility) and overall player performance is approximately linear, which explains the superiority of the Linear Regression model in this study.

Table 02: Feature Importance in the Random Forest Model

Ranking	Variable	Importance
1	Endurance	0.3643
2	Strength	0.3046
3	Agility	0.2105
4	Speed	0.1206

Source: Prepared by the researchers based on Python output

5.1. Feature Importance Analysis: The analysis of feature importance in the Random Forest model shows that Endurance is the most influential factor on physical performance with a proportion of (0.36), followed by Muscular Strength (0.30), then Agility (0.21), while Speed ranks last with (0.12).

It is also important to note that the physical performance of under-20 football players depends more on endurance and muscular strength compared to speed, which reflects the nature of continuous effort during matches rather than short explosive actions.

Overall, the results indicate that the linear model is more suitable for predicting physical performance in this sample. The studied physical variables explain a large portion of the variance in performance, with a clear dominance of endurance and muscular strength.

5.2. Model Evaluation Metrics: Evaluation metrics are used to measure the accuracy of predictive models and their ability to explain the relationship between independent variables and the dependent variable. Two main indicators were adopted in this study: the coefficient of determination (R^2) and the Root Mean Square Error (RMSE).

5.2.1. Coefficient of Determination (R^2): The coefficient of determination measures the model's ability to explain the variance in the dependent variable based on the independent variables.

$$\boldsymbol{R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}$$

Interpretation: The evaluation results of the predictive models using the coefficient of determination (R^2) show clear differences in explanatory power between the two models used in the study.

The Linear Regression model achieved a high R^2 value of 0.9025, indicating that the model is able to explain about 90% of the variance in players' physical performance. This reflects a strong linear relationship between the independent variables (speed, endurance, muscular strength, agility) and the dependent variable.

In contrast, the Random Forest model recorded a lower R^2 value of 0.7364, indicating weaker explanatory power compared to the linear model in this specific case.

Based on these findings, it can be concluded that the linear model performs better in prediction, suggesting that the relationship between variables is closer to a linear structure rather than a complex nonlinear one. These results also indicate that adding more variables is not necessary in this context, as the selected variables are sufficient to explain a large portion of performance variation.

Accordingly, based on the R^2 value, the Linear Regression model was adopted as the final predictive model due to its higher accuracy and stronger explanatory capability compared to the alternative model.

5.2.2. Root Mean Square Error (RMSE): RMSE measures the average deviation between actual and predicted values, and it expresses the magnitude of the error in the same unit as the original variable.

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}}$$

Interpretation: The evaluation results using the Root Mean Square Error (RMSE) show clear differences in predictive accuracy between the Linear Regression model and the Random Forest model. The Linear Regression model recorded an RMSE value of 2.24, which is relatively low, indicating that the average deviation between actual and predicted values is small. This reflects a high level of predictive accuracy and strong capability in estimating players' physical performance.

In contrast, the Random Forest model recorded a higher RMSE value of **3.69**, reflecting a greater prediction error and lower accuracy in representing the studied data. Based on these findings, the Linear Regression model proves to be superior in minimizing prediction error, which is consistent with the results of the R^2 metric and confirms that the relationship between the studied variables is closer to a simple linear structure.

Overall, the RMSE results confirm the superiority of the Linear Regression model over the Random Forest model, making it the most accurate and suitable model for predicting physical performance in this study.

Both evaluation metrics (R^2 and RMSE) were used to assess model performance. R^2 measures explanatory power, while RMSE measures the average prediction error. Higher R^2 values and lower RMSE values indicate better model performance and higher predictive accuracy. By combining both indicators, it becomes clear that:

- The Linear Regression model achieved the highest explanatory power (high R^2);
- It also achieved the lowest prediction error (low RMSE);
- Therefore, it is considered the best model for this study.

Conclusion: Based on the above findings, it can be concluded that integrating Python-based techniques in physical performance analysis has transformed sports evaluation for the U20 category from descriptive assessment to precise predictive analytics. The model demonstrated strong explanatory power (90%), confirming that physical variables are not isolated indicators but an interconnected system that can be modeled and optimized.

Moreover, the low error margin in predictions provides sports organizations with a reliable tool for forecasting player performance and designing individualized training programs based on accurate digital profiling. This opens the door to several practical implications and recommendations derived from hypothesis testing.

Hypothesis Testing: Based on both theoretical and applied results, the hypotheses are evaluated as follows:

First hypothesis: There is no statistically significant relationship between the physical variables (speed, strength, endurance, and agility) and the overall sports performance of players.

Result: The null hypothesis was rejected, and the alternative hypothesis was accepted. Pearson correlation analysis showed strong and positive relationships between the physical variables and performance, confirming that improving these components leads to an improvement in overall sports performance.
Second hypothesis: Player performance can be predicted using machine learning models such as Linear Regression and Random Forest.
Result: Accepted. The results demonstrated that predictive models can estimate performance with high accuracy, with Linear Regression outperforming Random Forest in terms of R^2 and RMSE.
Third hypothesis: There are no significant differences between traditional statistical models and machine learning models in predicting sports performance.
Result: Rejected. The results showed clear differences in favor of the Linear Regression model, which performed better than Random Forest in both accuracy and interpretability.
Study Results: The study reached a set of key findings, the most important of which are:
- There is a strong correlation between physical variables and overall player performance, according to Pearson Correlation analysis.
- The Linear Regression model outperformed the Random Forest model in terms of predictive performance.
- The Linear Regression model achieved a high R^2 value of approximately 0.90, indicating strong explanatory power of the independent variables.
- The same model recorded the lowest RMSE value of approximately 2.24, reflecting higher predictive accuracy;
- The results showed that Endurance and Strength are the most influential variables in overall physical performance.
- Linear models can be considered an effective tool for predicting sports performance in youth categories.

Recommendations: Based on the study results, the following recommendations can be made:
- Adopting statistical analysis and machine learning as supportive tools for evaluating sports performance within clubs;
- Focusing on developing endurance and muscular strength as they are the most influential factors in improving overall player performance;
- Using Python-based predictive models to continuously monitor player development;
- Expanding the sample size in future studies to obtain more accurate and generalizable results;
- Integrating more advanced data, such as GPS tracking and real match data, to improve predictive models;
- Conducting future comparisons with more advanced machine learning models such as XGBoost and Neural Networks;
- Encouraging sports clubs to establish digital databases for players to support modern sports analytics.

References List:
1. Alamar, K. (2013). Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers. Columbia: Columbia University Press.
2. Arnason, A., & et al. (2004). Physical fitness, injuries, and team performance in soccer. Medicine & Science in Sports & Exercise, 36(2).
3. Bangsbo, J., Iaia, F., & Krustrup, P. (2008). The Yo-Yo intermittent recovery test: a useful tool for evaluation of physical performance in intermittent sports. Sports Medicine, 38(1).
4. Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match outcomes in team sports: A review. Artificial Intelligence Review, 55(1).
5. Bush, M., & et al. (2015). Evolution of match performance parameters in the English Premier League. Journal of Human Kinetics, 45(1).

6. Helgerud, J., & et al. (2001). Aerobic endurance training improves soccer performance. Medicine & Science in Sports & Exercise, 33(11).
7. Herold, M., & et al. (2019). Machine learning in soccer: A systematic review. Science and Medicine in Football, 3(2).
8. Hoff, J., & Helgerud, J. (2004). Endurance and strength training for soccer players. Sports Medicine, 34(3).
9. Knudson, D. (2021). Fundamentals of Biomechanics. German-British academic publishing company: 3rd Edition, Springer Nature.
10. Little, T., & Williams, A. (2005). Specificity of acceleration, maximum speed, and agility in professional soccer players . Journal of Strength and Conditioning Research, 19(1).
11. Mohr, M., Krustrup, P., & Bangsbo, J. (2003). Match performance of high-standard soccer players with special reference to development of fatigue. Journal of Sports Sciences, 21(7).
12. Morgulev, E., Azar, O., & Lidor, R. (2018). Sports analytics and the big-data era: A survey. International Journal of Computer Science in Sport, 17(1).
13. Sheppard, J., & Young, W. (2006). Agility literature review: Classifications, training and testing. Journal of Sports Sciences, 24(9).
14. Stølen, T., & et al. (2005). Physiology of soccer: An update. Sports Medicine, 35(6).